The global brain semantic web : Interleaving human-machine knowledge and computation

Bernstein, Abraham

Abstract: Abstract : Before the Internet most collaborators had to be sufficiently close by to work together towards a certain goal. Now, the cost of collaborating with anybody anywhere on the world has been reduced to almost zero. As a result large-scale collaboration between humans and computers has become technically feasible. In these collaborative setups humans can carry the part of the weight of processing. Hence, people and computers become a kind of "global brain" of distributed interleaved human-machine computation (often called collective intelligence, social computing, or various other terms). Human computers as part of computational processes, however, come with their own strengths and issues.In this paper we take the underlying ideas of Bernstein et al. [1] regarding three traits on human computation—motivational diversity, cognitive diversity, and error diversity—and discuss them in the light of a Global Brain Semantic Web.

The Global Brain Semantic Web
—
Interleaving Human-Machine Knowledge and Computation*

Abraham Bernstein

Dynamic and Distributed Information Systems Group
University of Zurich, Switzerland
{lastname}@ifi.uzh.ch

Abstract. Before the Internet most collaborators had to be sufficiently close by to work together towards a certain goal. Now, the cost of collaborating with anybody anywhere on the world has been reduced to almost zero. As a result large-scale collaboration between humans and computers has become technically feasible. In these collaborative setups humans can carry the part of the weight of processing. Hence, people and computers become a kind of "global brain" of distributed interleaved human-machine computation (often called collective intelligence, social computing, or various other terms). Human computers as part of computational processes, however, come with their own strengths and issues.
In this paper we take the underlying ideas of Bernstein *et al.* [1] regarding three traits on human computation—motivational diversity, cognitive diversity, and error diversity—and discuss them in the light of a Global Brain Semantic Web.

Keywords: Semantic Web, Human Computation, Collective Intelligence, Crowdsourcing, Global Brain

1 Introduction

The Internet has changed the scale at which people and machines can collaborate. Before the Internet most collaborators had to be sufficiently close by to work together towards a certain goal. Now, the cost of collaborating with anybody anywhere on the world has been reduced to almost zero. As a consequence, new ways of combining networked humans and computers—whether they are called collective intelligence, social computing, or various other terms—are already important and likely to become truly transformative in domains from education and industry to government and the arts. These systems are now routinely able to solve problems that would have been unthinkably difficult only a few short

*This paper takes the underlying ideas of Bernstein *et al.* [1] and applies them to the Semantic Web setting. Hence, some of its content (especially Section 2) is derived from that publication. The application to the Semantic Web is novel.

years ago, combining the communication and number-crunching capabilities of computer systems with the creativity and high-level cognitive capabilities of people. As examples, just consider the wealth of information gathered by the Wikipedia,[1] the galaxies discovered by Galaxy Zoo,[2] or the reams of OCR tasks solved by ReCaptcha. [3] And all this is super-charged by the emergent generativity and robust evaluation that can come from the many eyes of large crowds of people. As the scale, scope, and connectivity of these human-computer networks increase, we believe it will become increasingly useful to view all the people and computers on our planet as constituting a kind of "global brain"—interconnected not by neurons but by the Internet.

The implication of using the capabilities of global brain computation on the Semantic Web (or Linked Data Could[4]) are manyfold. It has great potential to overcome some of the problems of the current Semantic Web. Queries with incomplete knowledge could be addressed as humans might guide additional data exploration, reasoning about information might be resolved by human intervention, labeling vast amounts of additional or the alignment of existing ontologies needed for answering a specific query could be achieved on the fly—these are but some examples for possibilities that such a Semantic Web might offer.

The current Semantic Web, however, is largely built on the vision of computers populating the Web of machines. Human computers [2] are different than machines. One one side, they need to motivated, they have a large variety of different capabilities, and are prone to commit an astonishingly large set of errors. On the other side, they can deal with tasks at which traditional computers simply fail. This gives rise to the question, how a Semantic Web would look like that incorporates human computers as first-class citizens. Said differently: How does the Global Brain Semantic Web—incorporating both human and machine computers and serving those two types—look like?

2 How are people different?[5]

As Bernstein *et al.* [1] write: "There are already literally hundreds of compelling examples of the global brain at work, collectively representing the contributions of many millions of people and computers [3][4]. These range from systems where individuals perform simple micro-tasks in Amazon's Mechanical Turk[6] to where they compete to solve complex engineering problems.[7]"

Their analysis continues: "While these systems cover an enormous range of approaches, it has become clear from these experiences that human computers

[1] `http://www.wikipedia.org`
[2] `http://www.galaxyzoo.org`
[3] `http://www.google.com/recaptcha`
[4] I will use these terms interchangeably in this text
[5] This section is essentially quoted from [1]. I changed some minor wording.
[6] `http://mturk.com`
[7] `innocentive.com`

are different than traditional computers in some fundamental ways. We list some of the most important differences below:

Motivational diversity People, unlike current computational systems, are self-interested and therefore require appropriate incentives—anything from money, fame, and fun to altruism and community—to perform tasks. These incentives have to be carefully designed, moreover, to avoid people gaming the system or causing outright damage. In some cases, one may even use their motivation to do one task to accomplish another, as in reCAPTCHA, where people OCR documents as a side effect of passing a human versus bot test.[5]

Cognitive diversity In most computer systems we deal with a limited range of diversity—in terms of memory, speed, and device access. People, by contrast, vary across many dimensions in the kinds of tasks they can do well, and their individual strengths are only incompletely understood at best. This implies qualitative differences in how (and how well) we can expect to match tasks and resources in a global brain context.

Error diversity With traditional computers, we worry much more about outright failure than other kinds of errors. And the other errors are usually highly deterministic and limited in diversity because a relatively small range of software is typically replicated across millions of computers. People, by contrast, are prone to a bewildering and inconsistent variety of idiosyncratic deviations from rational and accurate performance. The global brain, therefore, calls for a radically more capable quality assurance oriented towards the particular kinds of errors that occur with human participants. Fortunately, the global brain also provides access, at least currently, to a huge human "cognitive surplus" [6], so that, for instance, quality mechanisms based on previously-unthinkable levels of redundancy have become practical.

These attributes lead, in turn, to the possibility of new, and potentially troubling, forms of emergence. Crowds of people, when engaged in solving interdependent problems, can evince emergent behaviors that range from groupthink (where decision-makers converge prematurely on a small subset of the solution space) to balkanization (where decision-makers divide into intransigent competing cliques) to chaotic dynamics (e.g., stock market bubbles and crashes). While emergence is, of course, not unique to the global brain, it is probably made much more challenging by the unprecedented combination of microsecond computer and communications speeds, globe-scale interdependencies, and human diversity."

3 Why is the current Semantic Web is not adequate?

The diversities describe above give rise to a range of issues when including people into the Semantic Web. In some sense, we can see human computers play two roles in the context of the Semantic Web: they can be part of the "knowledge base" (i.e., the web of data) or they can be consumers of Semantic Web Data.

3.1 Humans as Part of the "Knowledge Base"

For centuries people have been part of knowledge bases. Librarians, for example, have fulfilled the role as "index mechanisms" to storages on knowledge, universities have served rulers as knowledge bases to be queried when need arose, or rumor mills have served as "social processing frameworks" to assess the likelihood of information and provide its reasoning capabilities to well-connected individuals. Hence, humans have been parts of knowledge bases for millennia. The current fabric of the Semantic Web is, however, not human-friendly.

First, the *motivation* of information sources has been completely ignored, as computers don't need to be motivated but programmed. Hence, the current framework of retrieving data via HTPP calls cannot be applied to people as they would have to be motivated to actually answer such a request. What we need is a motivational framework for human computers to participate in the Global Brain Semantic Web such as [7], where we propose to built an incentive structure for human computers that is economically well-founded. All these efforts are but the beginning of investigating the motivational incentives for human computers on the Semantic Web.

Second, the *cognitive capabilities* required in answering Semantic Web request (e.g., as linked data source, RDF-file server, or SPARQL endpoint) is well-structured for humans, but not suitable for the cognitive capabilities of humans. We need to develop protocols where human nodes in the web of data can be queried like question answering systems such as Yahoo! Answers[8] or human suitable tasks, as offered on Mechanical Turk.
In addition, we need to ask ourselves how the cognitive diversity can be exploited. Market mechanisms, for example, have served as a distributed mechanism to answer the question about the value of a good. Indeed, the market refines the differing knowledge of participants into a whole. As such markets exploit the cognitive diversity and varying degrees on information distribution for price determination. But what are those mechanisms for groups of people in a Semantic Web task? Some mechanisms [8][9] have been developed to go beyond simple averaging of opinions. But these explorations are at the beginning.

Third, we need to develop reasoning mechanisms that are robust against human *error diversity*. Sheng *et al.* [10], for example, ask themselves how many people should be asked to provide a label—translated into the Semantic Web content the value of a datatype property—in order to get a robust result. It turns out it is not always as many possible. Bernstein and Li [11] ask if the labeling quality of people can be improved if they are provided with some specific information. Others [8][9] provide intricate coordination mechanisms of how multiple people should be involved in ensuring the robustness of a system. Again, these findings but scratch on the surface of the problem of how to deal with error diversity.

[8] `http://answers.yahoo.com`

3.2 Humans as Consumers of Semantic Web Data

The provision of good human-interaction metaphors with the Semantic Web has received little attention in comparison to automated processing capabilities. Despite a series of ISWC and CHI workshops as well as a special Issue in the Journal of Web Semantics the design of Semantic Web user interactions is still an laborious undertaking. Interesting solutions exist for specific applications, as, for example, witnessed by some fascinating applications on the linked open data cloud.[9] The question of how people would interact with the enormous *generic* variety of data available on the Semantic Web is, however, largely unanswered and the generic Semantic Web Browser is still missing (and might never be available [12]). Note that part of the problem can be alleviated that we do not need a Semantic Web browser for the casual user (as defined by [13]), as human computers can be assumed to be motivated knowledge workers that are prepared to learn how to interact with the the Semantic Web just as librarians learn how to interact with large collections of documents. Nonetheless, if humans are to consume generic Semantic Web data we need novel approaches.

4 The Global Brain Semantic Web—A Call to Arms

The Global Brain Semantic Web—a Semantic Web interleaving a large number of human and machine computation—has great potential to overcome some of the issues of the current Semantic Web. To achieve these goals, however, we need to *address the challenges and opportunities offers by the three diversities* mentioned above and make *significant progress* along the lines of *interaction of educated users with Semantic Web-style data*.

New communities of research—be it in Web Science, Collective Intelligence, or Human Computation—are currently forming themselves. The Semantic Web research community needs to be bridge to these emergent activities in order to fulfill the vision of a robust and capable Semantic Web in the year 2012.

Acknowledgments

I am grateful for the contributions from and discussions with my co-authors Mark Klein and Tom Malone of the original Programming the Global Brain paper [1] as well as the insightful discussions with my graduate students and colleagues including Patrick Minder, Elaine Huang, Sven Seuken, Cosmin Basca, Esther Kaufmann, Katharian Reinecke, and Lorenz Fischer.

References

1. Bernstein, A., Klein, M., Malone, T.: Programming the global brain. Communications of the ACM **55**(5) (2012) 1–4

[9] see `https://explore.data.gov/catalog/apps/`, `http://data.gov.uk/apps`, or `http://make.opendata.ch/doku.php?id=event:home` (for a smaller, community-driven portal)

2. Grier, D.A.: When computers were human. Princeton University Press (2005)
3. Doan, A., Ramakrishnan, R., Halevy, A.Y.: Crowdsourcing systems on the world-wide web. Commun. ACM **54**(4) (April 2011) 86–96
4. Quinn, A., Bederson, B.: Human computation: a survey and taxonomy of a growing field. In: Proceedings of the 2011 annual conference on Human factors in computing systems, ACM (2011) 1403–1412
5. Von Ahn, L., Maurer, B., McMillen, C., Abraham, D., Blum, M.: recaptcha: Human-based character recognition via web security measures. Science **321**(5895) (2008) 1465
6. Shirky, C.: Cognitive Surplus: Creativity and Generosity in a Connected Age. Penguin Press (10§0)
7. Minder, P., Seuken, S., Bernstein, A., Zollinger, M.: Crowdmanager - combinatorial allocation and pricing of crowdsourcing tasks with time constraints. In: Workshop on Social Computing and User Generated Content in conjunction with ACM Conference on Electronic Commerce (ACM-EC 2012), Valencia, Spain (JUN 2012) 1–18
8. Bernstein, M., Little, G., Miller, R., Hartmann, B., Ackerman, M., Karger, D., Crowell, D., Panovich, K.: Soylent: a word processor with a crowd inside. In: Proceedings of the 23nd annual ACM symposium on User interface software and technology, ACM (2010) 313–322
9. Minder, P., Bernstein, A.: How to translate a book within an hour - towards general purpose programmable human computers with crowdlang. In: Web Science 2012, New Yortk, NY, USA (JUN 2012)
10. Sheng, V.S., Provost, F., Ipeirotis, P.G.: Get another label? improving data quality and data mining using multiple, noisy labelers. In: Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. KDD '08, New York, NY, USA, ACM (2008) 614–622
11. Bernstein, A., Li, J.: From active towards interactive learning: using consideration information to improve labeling correctness. In: Human Computation Workshop. (2009)
12. Karger, D., Schraefel, M.: The pathetic fallacy of rdf (2006)
13. Kaufmann, E., Bernstein, A.: Evaluating the usability of natural language query languages and interfaces to semantic web knowledge bases. Journal of Web Semantics **8**(4) (2010) 377–393